Research Data Management for Reproducibility

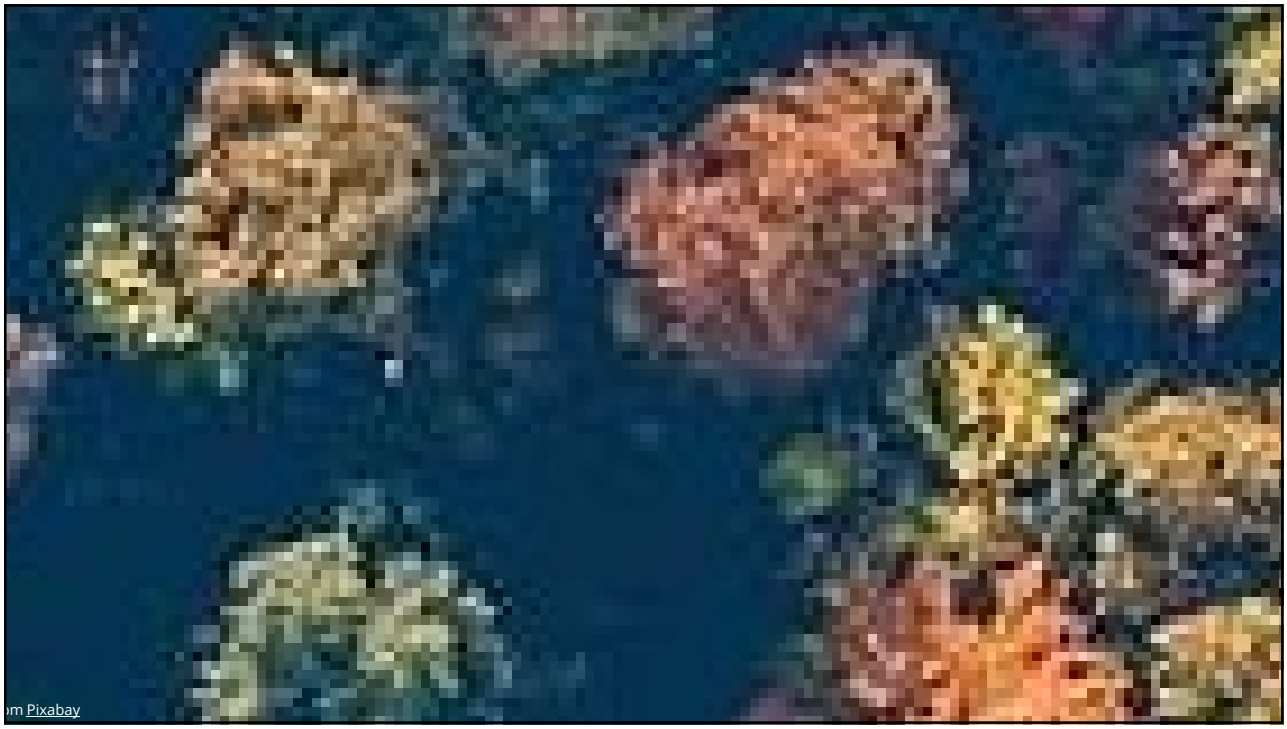
Compute Ontario Colloquium Series

February 7, 2024

Jeff Moon
Director, Data Strategy & Services
Compute Ontario
moonj@computeontario.ca

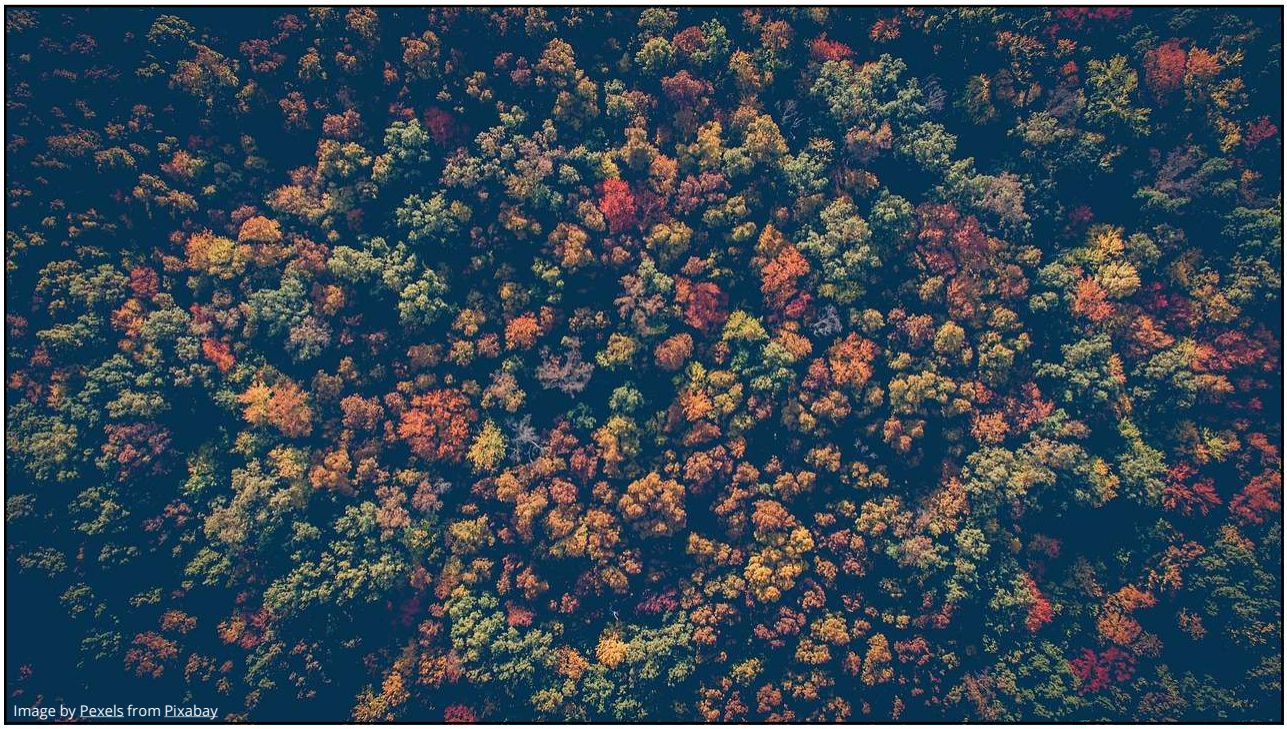
Image by Gerd Altmann from Pixabay

1



m Pixabay

2

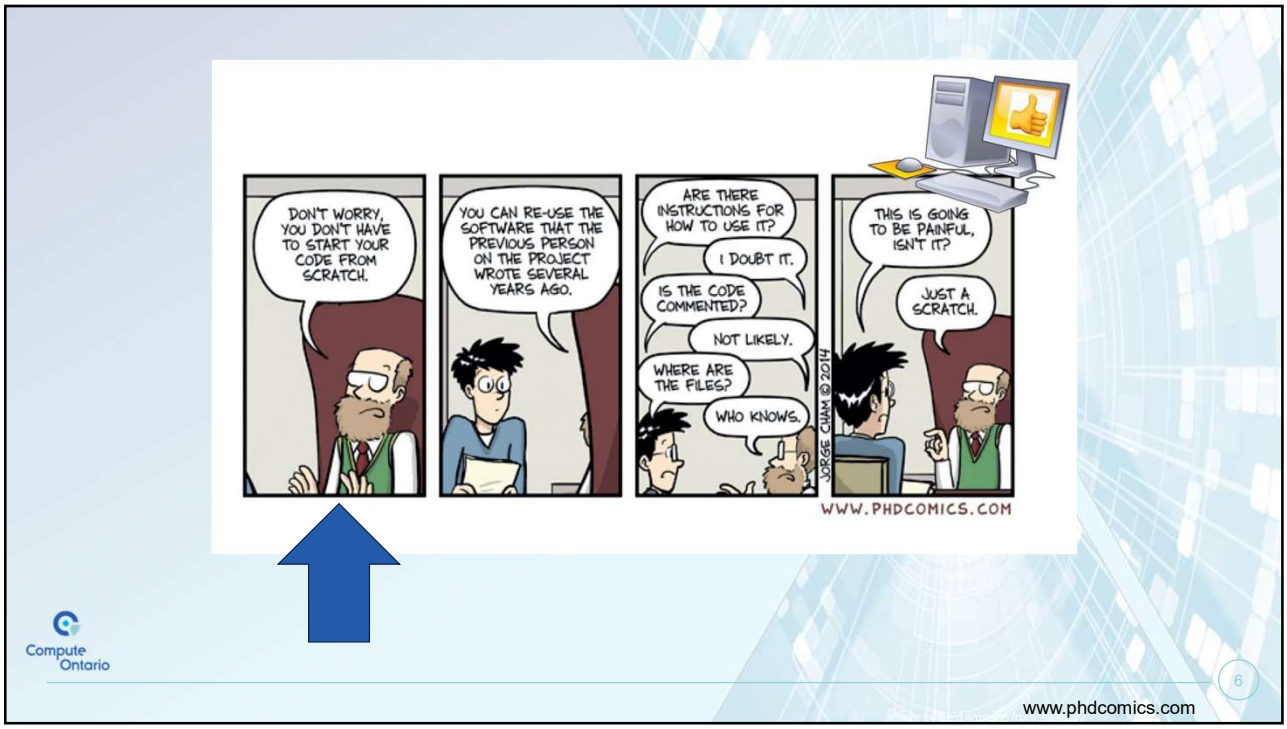


3

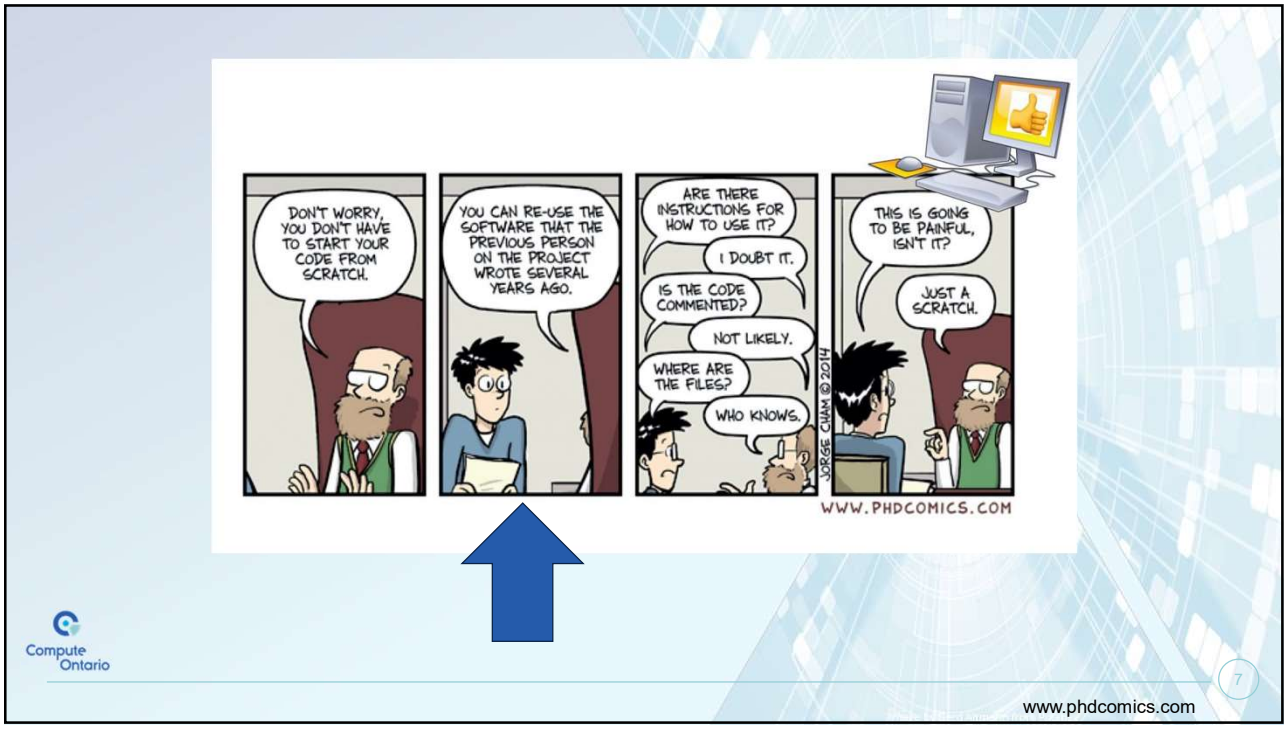
What we'll cover today:

1. Some background
2. Exploring the concept of reproducibility
3. Challenges & benefits
4. Applying FAIR principles
5. Possible policy directions

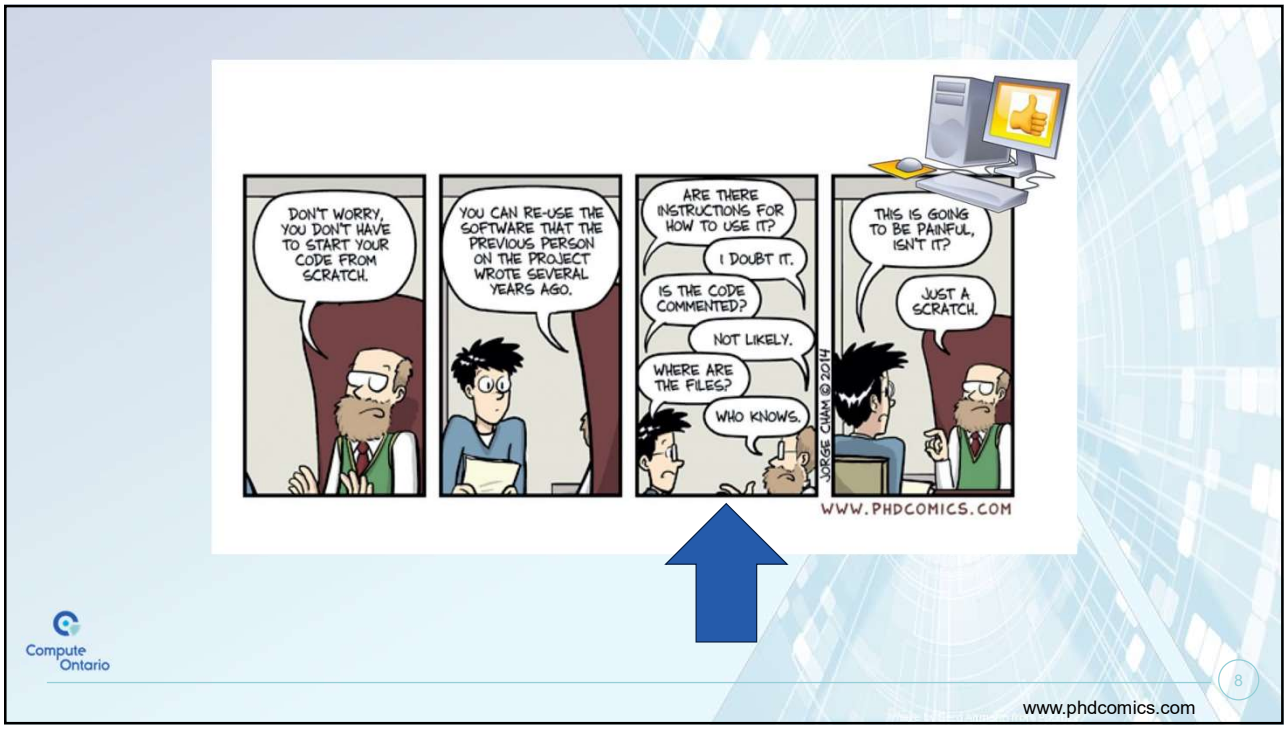
4


DON'T WORRY, YOU DON'T HAVE TO START YOUR CODE FROM SCRATCH.
YOU CAN RE-USE THE SOFTWARE THAT THE PREVIOUS PERSON ON THE PROJECT WROTE SEVERAL YEARS AGO.
ARE THERE INSTRUCTIONS FOR HOW TO USE IT?
I DOUBT IT.
IS THE CODE COMMENTED?
NOT LIKELY.
WHERE ARE THE FILES?
WHO KNOWS.
JORGE CHAM © 2014
THIS IS GOING TO BE PAINFUL, ISN'T IT?
JUST A SCRATCH.
WWW.PHDCOMICS.COM
Compute Ontario
6
www.phdcomics.com

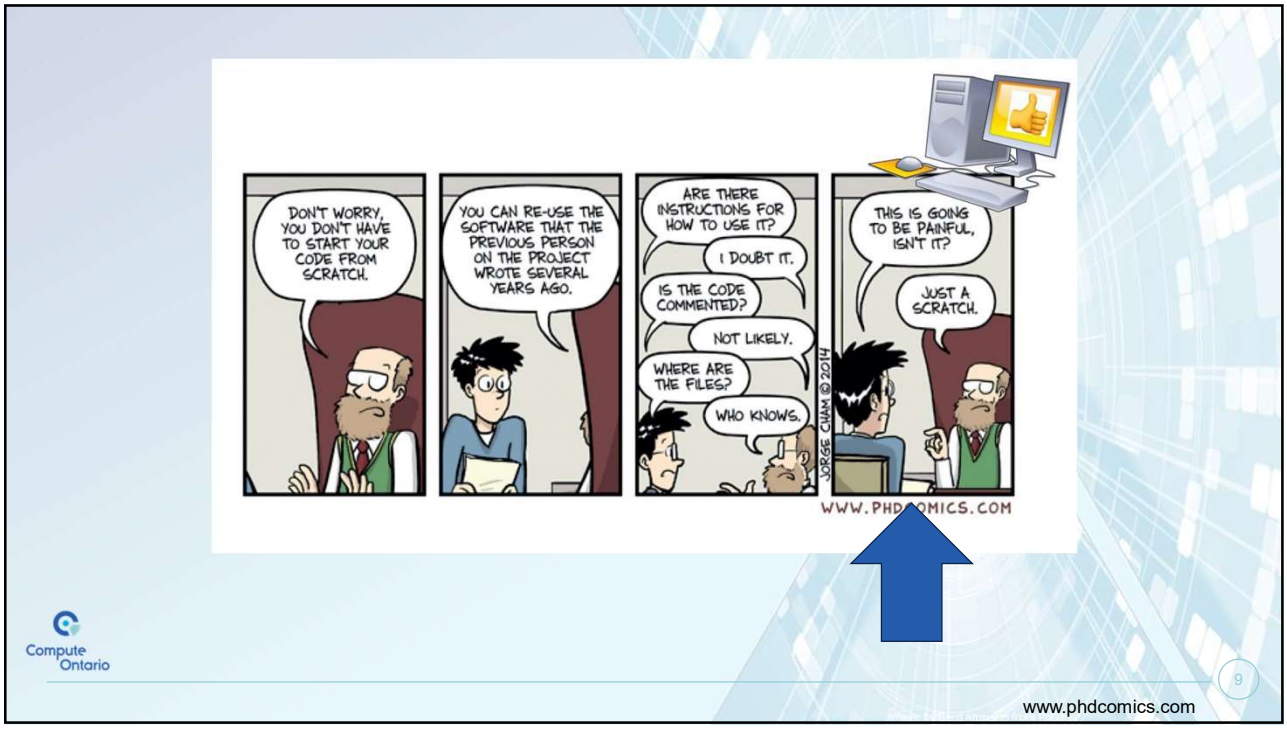
6


DON'T WORRY, YOU DON'T HAVE TO START YOUR CODE FROM SCRATCH.
YOU CAN RE-USE THE SOFTWARE THAT THE PREVIOUS PERSON ON THE PROJECT WROTE SEVERAL YEARS AGO.
ARE THERE INSTRUCTIONS FOR HOW TO USE IT?
I DOUBT IT.
IS THE CODE COMMENTED?
NOT LIKELY.
WHERE ARE THE FILES?
WHO KNOWS.
JORGE CHAM © 2014
THIS IS GOING TO BE PAINFUL, ISN'T IT?
JUST A SCRATCH.
WWW.PHDCOMICS.COM
Compute Ontario
7
www.phdcomics.com

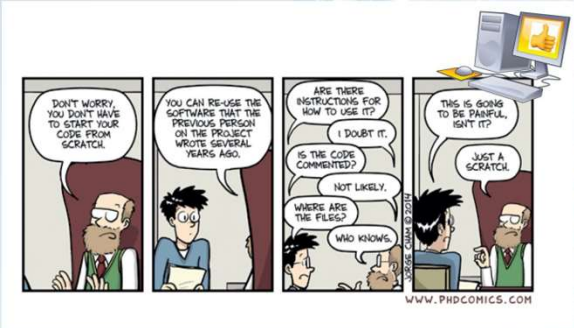
7


DON'T WORRY, YOU DON'T HAVE TO START YOUR CODE FROM SCRATCH.
YOU CAN RE-USE THE SOFTWARE THAT THE PREVIOUS PERSON ON THE PROJECT WROTE SEVERAL YEARS AGO.
ARE THERE INSTRUCTIONS FOR HOW TO USE IT?
I DOUBT IT.
IS THE CODE COMMENTED?
NOT LIKELY.
WHERE ARE THE FILES?
WHO KNOWS.
JORGE CHAM © 2014
THIS IS GOING TO BE PAINFUL, ISN'T IT?
JUST A SCRATCH.
WWW.PHDCOMICS.COM
Compute Ontario
8
www.phdcomics.com

8


DON'T WORRY, YOU DON'T HAVE TO START YOUR CODE FROM SCRATCH.
YOU CAN RE-USE THE SOFTWARE THAT THE PREVIOUS PERSON ON THE PROJECT WROTE SEVERAL YEARS AGO.
ARE THERE INSTRUCTIONS FOR HOW TO USE IT?
I DOUBT IT.
IS THE CODE COMMENTED?
NOT LIKELY.
WHERE ARE THE FILES?
WHO KNOWS.
JORGE CHAM © 2014
THIS IS GOING TO BE PAINFUL, ISN'T IT?
JUST A SCRATCH.
WWW.PHDCOMICS.COM
Compute Ontario
9
www.phdcomics.com

9



A key goal of **Research Data Management** is to make reuse of data, code, and documentation as painless as possible

Compute Ontario

10

www.phdcomics.com

10

Evolving research practices

HPC

Compute Ontario

Miriam Martínez - 16 MARCH 2022 BLOG POST: https://www.zeclinics.com/blog/differences-between-in-vitro-in-vivo-and-in-silico-assays-in-preclinical-research/ Broude Geva, S. et al. **Fostering Collaboration Among Organizations in the Research Computing and Data Ecosystem,** Practice & Experience in Advanced Research Computing (PEARC) Conference Series, July 26–30, 2020, Portland, OR, USA

11

11

Reproducibility

"…obtaining consistent results using the same input data, computational steps, methods, code, and conditions of analysis"

National Academies of Sciences, E. (2019). *Reproducibility and replicability in science*. Washington, District of Columbia: National Academies Press. https://nap.nationalacademies.org/catalog/25303/reproducibility-and-replicability-in-science p. 46

12

12

Reproducibility

Computational Chemistry

Health

Ecology

Genomics

Hydrology

AI & Machine Learning

Natural Language Processing

Astronomy

Clinical Metabolomics

+ more…

Leipzig, Nüst, D., Hoyt, C. T., Ram, K., & Greenberg, J. (2021). The role of metadata in reproducible computational research. Patterns (New York, N.Y.), 2(9), 100322–100322. https://doi.org/10.1016/j.patter.2021.100322

13

13

1990's

Jon Claerbout

"A revolution in education and technology transfer follows from the **marriage of word processing and software command scripts**. In this marriage an author attaches to every figure caption a **pushbutton** or name tag usable *to recalculate the figure from all its data, parameters, and programs*"

And went on to perhaps rather naively state that:

"preparing such electronic documents *is little effort beyond our customary report writing*; **mainly we need to file everything in a systematic way**"

National Academies of Sciences, E. (2019). *Reproducibility and replicability in science*. Washington, District of Columbia: National Academies Press. https://nap.nationalacademies.org/catalog/25303/reproducibility-and-replicability-in-science

Claerbout, J. F. and M. Karrenbach, 1992: Electronic documents give reproducible research a new meaning. In *SEG Technical Program Expanded Abstracts 1992*, Society of Exploration Geophysicists, pp. 601–604, doi:10.1190/1.1822162.

Image by callum ramsay from Pixabay

14

14

They went on to build a CD-ROM based resource to:

- Merge a publication with its underlying computational analysis
- Preserve the local software environment
- Provide 'push button' recalculation of results
- Merge and link multiple electronic documents
- Export documents to facilitate reproduction by others

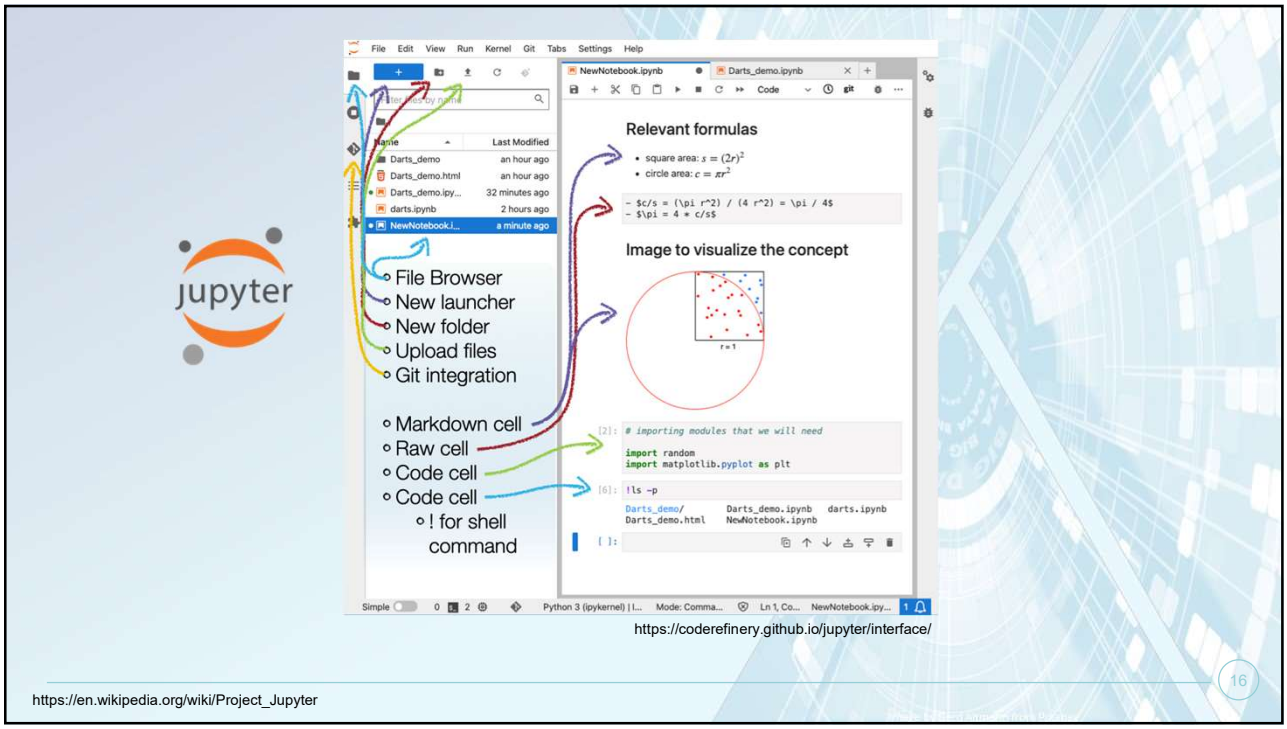
"The CD-ROM, at 680 megabytes, is **so large** we have had room for many executable programs on popular brands of workstations"

Claerbout, J. F. and M. Karrenbach, 1992: Electronic documents give reproducible research a new meaning. In *SEG Technical Program Expanded Abstracts 1992*, Society of Exploration Geophysicists, pp. 601–604, doi:10.1190/1.1822162.

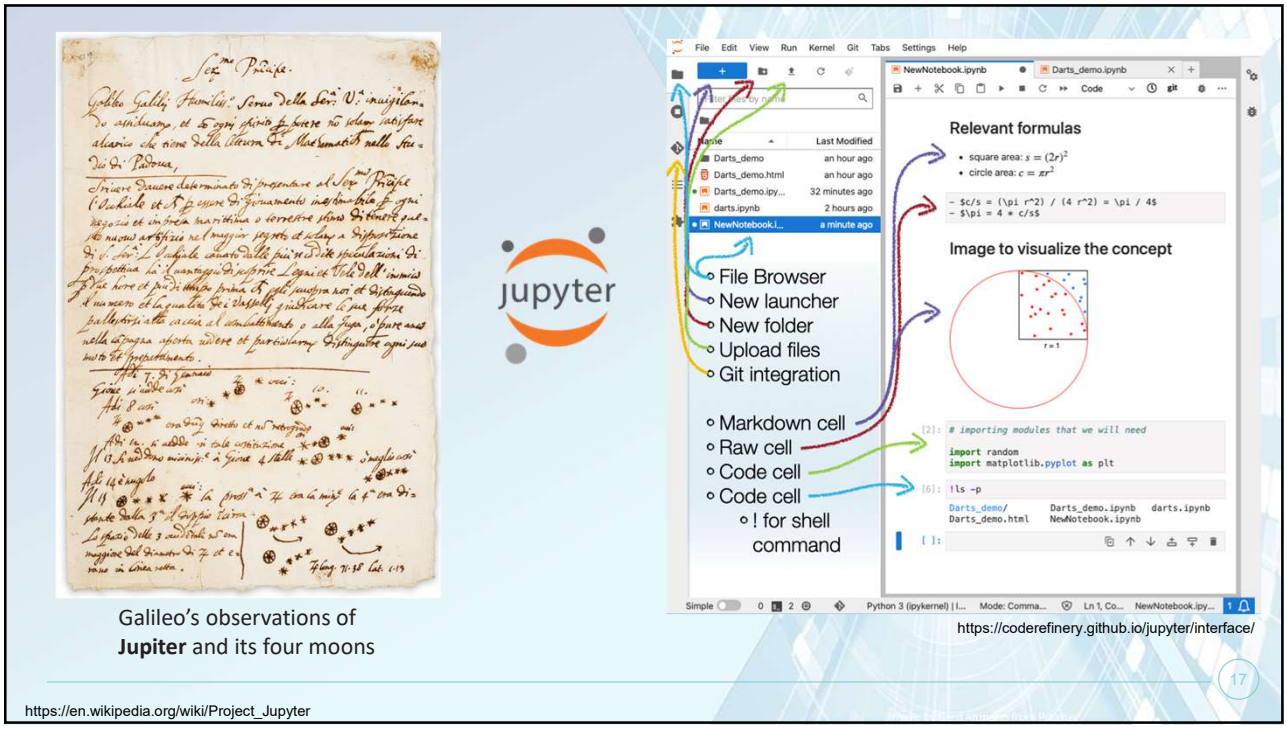
Image from: Wikipedia

15

15


File Browser
New launcher
New folder
Upload files
Git integration
Markdown cell
Raw cell
Code cell
Code cell
! for shell command
Relevant formulas
Image to visualize the concept
https://coderefinery.github.io/jupyter/interface/
https://en.wikipedia.org/wiki/Project_Jupyter
16

16


Galileo's observations of
Jupiter and its four moons
File Browser
New launcher
New folder
Upload files
Git integration
Markdown cell
Raw cell
Code cell
Code cell
! for shell command
Relevant formulas
Image to visualize the concept
https://coderefinery.github.io/jupyter/interface/
https://en.wikipedia.org/wiki/Project_Jupyter
17

17

David Donoho, *et al*

*"An **article** about computational science in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship".*

*"The **actual scholarship** is the complete software development environment and… instructions which generated the figures."*

Curating Data Sets for Reproducibility Workshop; Q. Zhang, S. Sawchuk, S. Khair, https://research-reuse.github.io
Barba, L. A. (2018). Terminologies for Reproducible Research. *ArXiv.Org*. https://doi.org/10.48550/arxiv.1802.03311
Claerbout, J. F. and M. Karrenbach, 1992: Electronic documents give reproducible research a new meaning. In *SEG Technical Program Expanded Abstracts 1992*, Society of Exploration Geophysicists, pp. 601–604, doi:10.1190/1.1822162.

18

Reproducibility

Code	Data: Same	Data: Different
Same	Reproducible	Replicable
Different	Robust	Generalizable

Figure 1. Whitaker's matrix of reproducibility
Whitaker's matrix of reproducibility;[10] made available under the Creative Commons Attribution license (CC-BY 4.0).

"bit-reproducibility"

Reproducible: Research is reproducible if we can re-run an experiment using the same method (Code) in the same environment (HPC) using the same data and obtain the same results.

"conclusion-reproducibility"

Replicable: If the underlying scientific hypothesis can be independently confirmed, post-publication, using the same method (i.e. code) but different data.

Robust: If different code using the same data supports the same conclusions.

Generalizable: If different code and different data can be used to support the same conclusions.

Sources and/or derived from:

Melsen, L. A., Torfs, P. J. J. ., Uijlenhoet, R., & Teuling, R. (2017). Comment on "Most computational hydrology is not reproducible, so is it really science?" by Christopher Hutton et al. *Water Resources Research*, *53*(3), 2568–2569. https://doi.org/10.1002/2016WR020208
Collberg, & Proebsting, T. (2016). Repeatability in computer systems research. Communications of the ACM, 59(3), 62–69. https://doi.org/10.1145/2812803
Whitaker, K. (2016). Showing your working: a guide to reproducible neuroimaging analyses. Figshare. https://doi.org/10.6084/m9.figshare.4244996.v1.
Leipzig, Nüst, D., Hoyt, C. T., Ram, K., & Greenberg, J. (2021). The role of metadata in reproducible computational research. Patterns (New York, N.Y.), 2(9), 100322–100322. https://doi.org/10.1016/j.patter.2021.100322

19

19

Reproducibility

Code	Data: Same	Data: Different
Same	Reproducible	Replicable
Different	Replicable	Replicable

Figure 1. Whitaker's matrix of reproducibility
Whitaker's matrix of reproducibility;[10] made available under the Creative Commons Attribution license (CC-BY 4.0).

Reproducible

same data + same methods
= same ***results***

Replicable

new data *and/or* new methods
in an independent study
= same ***findings***

Sources and/or derived from:

Melsen, L. A., Torfs, P. J. J. ., Uijlenhoet, R., & Teuling, R. (2017). Comment on "Most computational hydrology is not reproducible, so is it really science?" by Christopher Hutton et al. *Water Resources Research*, *53*(3), 2568–2569. https://doi.org/10.1002/2016WR020208
Collberg, & Proebsting, T. (2016). Repeatability in computer systems research. Communications of the ACM, 59(3), 62–69. https://doi.org/10.1145/2812803
Whitaker, K. (2016). Showing your working: a guide to reproducible neuroimaging analyses. Figshare. https://doi.org/10.6084/m9.figshare.4244996.v1.
Leipzig, Nüst, D., Hoyt, C. T., Ram, K., & Greenberg, J. (2021). The role of metadata in reproducible computational research. Patterns (New York, N.Y.), 2(9), 100322–100322. https://doi.org/10.1016/j.patter.2021.100322

20

20

How these terms are used in practice

A
No distinction

Reproducible
Replicable

B
treated distinctly

B1
Reproducible
same data + same methods
= same results

B2
Replicable
new data *and/or* new methods in an independent study
= same findings

Barba, L. A. (2018). Terminologies for Reproducible Research. *ArXiv.Org*. https://doi.org/10.48550/arxiv.1802.03311

21

21

Evidence from the Literature

A No distinction

Reproducible

Replicable

King (1995), 527
JCGM (2008), 32

Dewald et al. (1986), 506
Pesaran (2003), 12
McCullough et al. (2008), 93
Garijo et al. (2013), 52
Open Science Collaboration (2012), 300
Open Science Collaboration (2015), 1573
Stodden (2015), 19
Duvendack et al. (2017), 13
Lejaeghere et al. (2016),199
Coudert (2017), 3

B treated distinctly

B1

Reproducible

same data + same methods = same results

Peng et al. (2006), 177
Gentleman and Temple Lang (2007), 216
Laine et al. (2007), 134
Vandewalle et al. (2009), 266
LeVeque (2009), 32
Hyndman (2010), 20
Jasny et al. (2011), 180
Peng (2011), 552

Koenker and Zeileis (2009), 58
Delescluse et al. (2012), 22
Sandve et al. (2013), 227
Stodden et al. (2014), 119
Topalidou et al. (2015), 14
Iqbal et al. (2016), 67
Kafkafi et al. (2016), 2
Stevens (2017), 1
Kitzes et al. (2017), 10
Benureau and Rougier (2017), 1
Bollen et al. (2015), 12
Broman et al. (2017), 4

B2

Replicable

new data *and/or* new methods in an independent study = same findings

Drummond (2009), 135
Casadevall and Fang (2010), 58
Stodden (2011), 30
Davison (2012), 80
Loscalzo (2012), 31
LeVeque et al. (2012), 74
Crook et al. (2013), 16
Cooper et al. (2015), 26

Cartwright (1991), 81
Pellizzari et al. (2017)
FASEB (2016)

Barba, L. A. (2018). Terminologies for Reproducible Research. *ArXiv.Org*. https://doi.org/10.48550/arxiv.1802.03311

22

Usage Grouped by Discipline

A No distinction

Reproducible

Replicable

political science
economics

B treated distinctly

B1

Reproducible

same data + same methods = same results

signal processing
scientific computing
econometry
epidemiology
clinical studies
internal medicine
physiology (neuro)
computational biology
biomedical research
statistics

B2

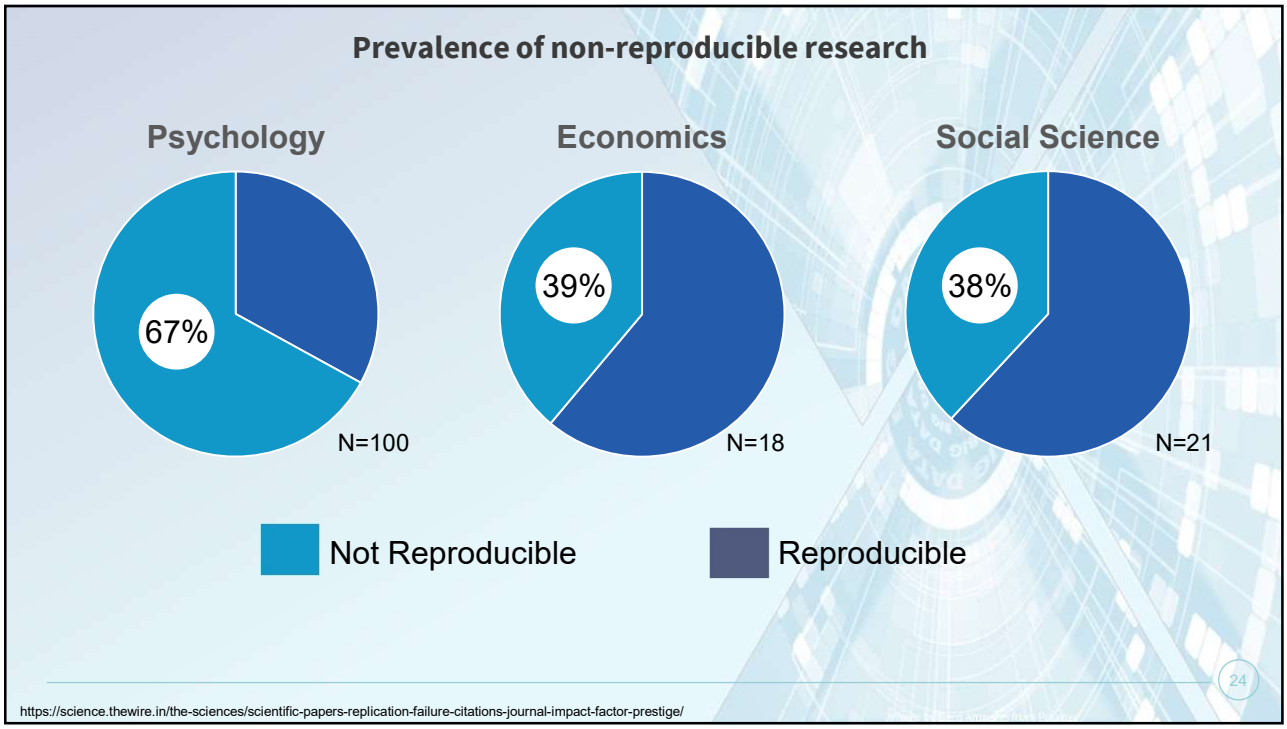
Replicable

new data *and/or* new methods in an independent study = same findings

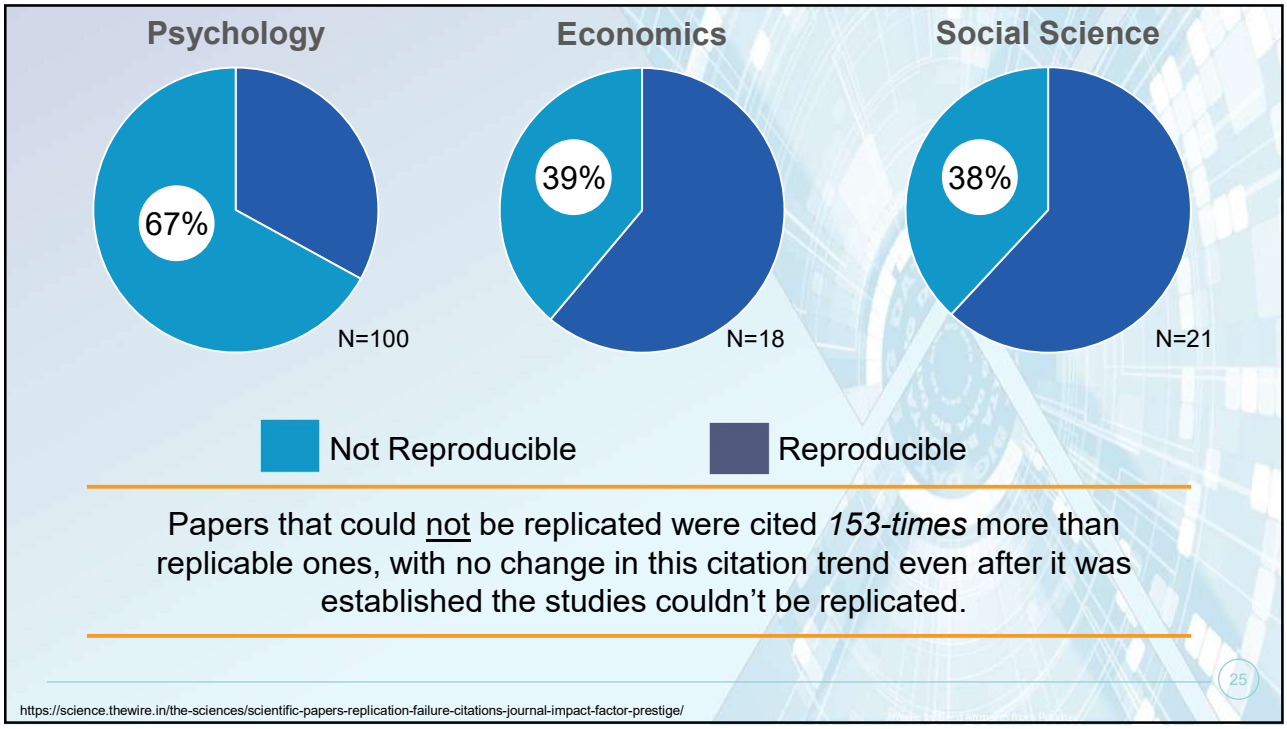
microbiology, immunology
computer science

Barba, L. A. (2018). Terminologies for Reproducible Research. *ArXiv.Org*. https://doi.org/10.48550/arxiv.1802.03311

23


Prevalence of non-reproducible research
Psychology
Economics
Social Science
67%
39%
38%
N=100
N=18
N=21
Not Reproducible
Reproducible
24
https://science.thewire.in/the-sciences/scientific-papers-replication-failure-citations-journal-impact-factor-prestige/

24


Psychology
Economics
Social Science
67%
39%
38%
N=100
N=18
N=21
Not Reproducible
Reproducible
Papers that could not be replicated were cited 153-times more than replicable ones, with no change in this citation trend even after it was established the studies couldn't be replicated.
25
https://science.thewire.in/the-sciences/scientific-papers-replication-failure-citations-journal-impact-factor-prestige/

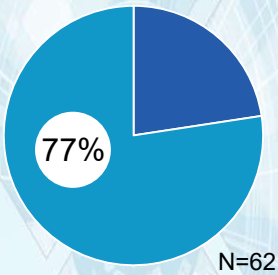
25

Prevalence of non-reproducible research

Computer Science

Unable to get code up & running within 30 minutes

N=402



N=62

Not Reproducible

Reproducible

Software as research data: Tips for Reproducible Research based on Software.
https://zenodo.org/record/2611314/files/Software-Manage-Publish-EN.pptx?download=1

Collberg Christian, Proebsting, Todd and Warren, Alex M. Repeatability and Benefaction in Computer Systems Research. , studie (2015).
McGeary, Kerry Anne, McCullough, B. D. and Harrison, Teresa. "Lessons from the JMCB archive." *Journal of money credit and banking* 38 , no. 4 (2006): 1093-1107.

26

26

HOW TO WRITE A PAPER IN 40 STEPS

MAKE DRAFT	STEP 1	Make a working **title**
	STEP 2	Introduce the **topic** and define terminology
	STEP 3	Emphasize why is the topic important
	STEP 4	Relate to **current knowledge**: what's been done
	STEP 5	Indicate the **gap**: what need's to be done?
	STEP 6	Pose research **questions**
	STEP 7	Give purpose and **objectives**
	STEP 8	List methodological **steps**
	STEP 9	Explain **theory** behind the methodology used
	STEP 10	Describe **experimental set-up**
	STEP 11	Describe **object** of the study (technical details)
	STEP 12	Give summary **results**
	STEP 13	Compare different results
	STEP 14	Focus on main **discoveries**
	STEP 15	Answer research questions (**conclusions**)
	STEP 16	Support and defend **answers**
	STEP 17	Explain conflicting results, unexpected findings and **discrepancies** with other research
	STEP 18	State **limitations** of the study
	STEP 19	State importance of findings
	STEP 20	Establish **newness**
	STEP 21	Announce **further research**
	STEP 22	ABSTRACT: what was done, what was found and what are the main conclusions
REVISE	STEP 23	Is the title clear and does it reflect the content and main findings?
	STEP 24	Are key terms clear and familiar?
	STEP 25	Are the objectives clear and relevant to the audience?
	STEP 26	Are all variables, techniques and materials listed, explained and linked to existing knowledge - are the results reproducible?
	STEP 27	Are all results and comparisons relevant to the posed questions/objectives?
	STEP 28	Do some statements and findings repeat in the text, tables of figures?
	STEP 29	Do the main conclusions reflect the posed questions?
	STEP 30	Will the main findings be unacceptable by the scientific community?
	STEP 31	Is the text coherent, clear and focused on a specific problem/topic?
	STEP 32	Is the abstract readable standalone (does it reflects the main story)?
POLISH	STEP 33	Are proper tenses and voices used (active and passive)?
	STEP 34	Are all equations mathematically correct and explained in the text?
	STEP 35	Are all abbreviations explained?
	STEP 36	Reconsider (avoid) using of words "very", "better", "may", "appears", "more", "convinced", "impression" in the text.
	STEP 37	Are all abbreviations, measurement units, variables and techniques internationally recognised (IS)?
	STEP 38	Are all figures/tables relevant and of good quality?
	STEP 39	Are all figures, tables and equations listed and mentioned in the text?
	STEP 40	Are all references relevant, up to date and accessible?

Source: Hengl, T. and Gould, M., 2002. Rules of thumb for writing research articles.

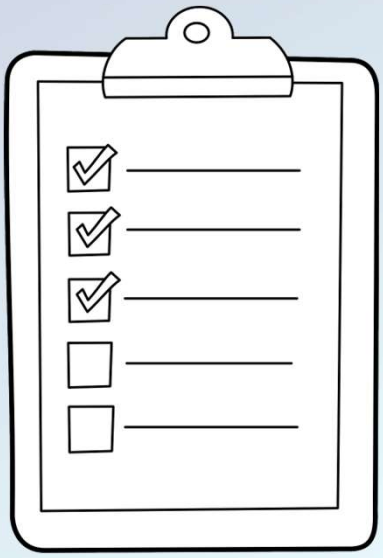
2002 checklist for writing a research article

Makes no mention of managing and providing access to data, metadata, or code that underpins research
&
supports reproducibility

Hengl, Tomislav, and Michael Gould. "Rules of thumb for writing research articles." *Enschede, September* (2002). https://webapps.itc.utwente.nl/librarywww/papers/hengl_rules.pdf

27

27



Reproducibility Checklists

Du, X., Aristizabal-Henao, *et al*, 2022. **A Checklist for Reproducible Computational Analysis in Clinical Metabolomics Research**. Metabolites, 12(1), 87–. https://doi.org/10.3390/metabo12010087

Hutton, C., *et al*, 2016. **Most computational hydrology is not reproducible, so is it really science?** Water Resources Research, 52(10), 7548–7555. https://doi.org/10.1002/2016WR019285

Sandve, Geir Kjetil, *et al*. **Ten Simple Rules for Reproducible Computational Research.** PLOS Computational Biology 9 , no. 10 (2013): 1-4.

Software as research data: Tips for Reproducible Research based on Software. https://zenodo.org/record/2611314/files/Software-Manage-Publish-EN.pptx?download=1

British Ecological Society. A Guide to Reproducible Code in Ecology and Evolution, 2017. https://www.britishecologicalsociety.org/wp-content/uploads/2017/12/guide-to-reproducible-code.pdf

Joelle Pineau, 2020. **The Machine Learning Reproducibility Checklist**. https://www.cs.mcgill.ca/~jpineau/ReproducibilityChecklist.pdf

James, Wilkins-Diehr, *et al*, (2014). **Standing Together for Reproducibility in Large-Scale Computing: Report on reproducibility@XSEDE**. https://doi.org/10.48550/arxiv.1412.5557

Crick, Hall, B. A., & Ishtiaq, S. (2017). **Reproducibility in Research: Systems, Infrastructure, Culture**. Journal of Open Research Software, 5(1), 32–. https://doi.org/10.5334/jors.73

Compute Ontario

28

Failure to replicate can occur for a number of reasons, including:

1. The first study's methods were flawed	and no relationship exists between variables
2. The second study's methods were flawed	so did not confirm a true relationship between variables
3. The two studies may, in fact, align but…	sampling variation might mask statistical significance in the second study
4. The methods/conditions in the second study were different	a mismatch in key elements needed for replication challenges related to lack of best practices for replication in the original study

Social, Behavioral, and Economic Sciences Perspectives on Robust and Reliable Science - Report of the Subcommittee on Replicability in Science Advisory Committee to the National Science Foundation Directorate for Social, Behavioral, and Economic Sciences, May 2015 https://www.nsf.gov/sbe/AC_Materials/SBE_Robust_and_Reliable_Research_Report.pdf

29

For HPC, these challenges include:

the cost of repeating computationally-intensive research is high

HPC resources are allocated competitively, discouraging replication

processes run on different computers can yield different results

code changes and data post-processing steps may be poorly documented (Medina, 2022)

data and/or software can be proprietary or otherwise restricted

software stacks evolve quickly

HPC systems are decommissioned every few years

software that supports reproducibility may not perform as well as proprietary alternatives (Courtes, 2022)

Compute Ontario

Medina, J. et al (2022). Accelerating the adoption of research data management strategies. Matter, Volume 5, Issue 11, 2 November 2022, Pages 3614-3642 https://doi.org/10.1016/j.matt.2022.10.007
Plale, Malik, T., Pouchard, L. C., Barba, L. A., & Gesing, S. (2021). Reproducibility Practice in High-Performance Computing: Community Survey Results. *Computing in Science & Engineering*, *23*(5), 55–60. https://doi.org/10.1109/MCSE.2021.3096678
Courtes. (2022). Reproducibility and Performance: Why Choose? Computing in Science & Engineering, 24(3), 77–80. https://doi.org/10.1109/MCSE.2022.3165626
Social, Behavioral, and Economic Sciences Perspectives on Robust and Reliable Science - Report of the Subcommittee on Replicability in Science Advisory Committee to the National Science Foundation Directorate for Social, Behavioral, and Economic Sciences, May 2015 https://www.nsf.gov/sbe/AC_Materials/SBE_Robust_and_Reliable_Research_Report.pdf

30

But there are real benefits to overcoming these challenges

Increased transparency

Verifying & building upon reported findings

Improving research methods

Preserving a complete scientific record

Complying with journal & funder policies

Improved training

Enhancing reputation of research & researchers

Reducing duplication

Compute Ontario

10 Things for Curating Reproducible and FAIR Research [RDA Recommendation] Florio Arguillas, Thu-Mai Christian, Mandy Gooch, Tom Honeyman, Limor Peer, CURE-FAIR WG; 27 June 2022; DOI: 10.15497/RDA00074 or https://zenodo.org/records/6797657#.YuB6zcHMKrd

31

RESEARCH DATA ALLIANCE

Challenges of Curating for Reproducible and FAIR Research Output

A Report by the RDA CURE-FAIR Working Group, Subgroup 3 on CURE-FAIR Challenges

Limor Peer, Florio Arguillas, Tom Honeyman, Nadica Miljković, Karsten Peters-von Gehlen and CURE-FAIR subgroup 3

April 12, 2021

Computational reproducibility **is the ability to obtain consistent computational results** using the same input data, computational steps, methods, code, and conditions of analysis.

Image by Mohamed Hassan from Pixabay

32

Peer, L., Arguillas, F., Honeyman, T., Miljković, N., Peters-von-Gehlen, K., & CURE-FAIR WG Subgroup 3. (2021). Challenges of Curating for Reproducible and FAIR Research Output. *Research Data Alliance*. DOI: 10.15497/RDA00063 https://www.rd-alliance.org/sites/default/files/Challenges%20of%20Curating%20for%20Reproducible%20and%20FAIR%20Research%20Output%20-%20Output%20Card_0.pdf

32

RESEARCH DATA ALLIANCE

Challenges of Curating for Reproducible and FAIR Research Output

A Report by the RDA CURE-FAIR Working Group, Subgroup 3 on CURE-FAIR Challenges

Limor Peer, Florio Arguillas, Tom Honeyman, Nadica Miljković, Karsten Peters-von Gehlen and CURE-FAIR subgroup 3

April 12, 2021

Computational reproducibility **is the ability to obtain consistent computational results** using the same input data, computational steps, methods, code, and conditions of analysis.

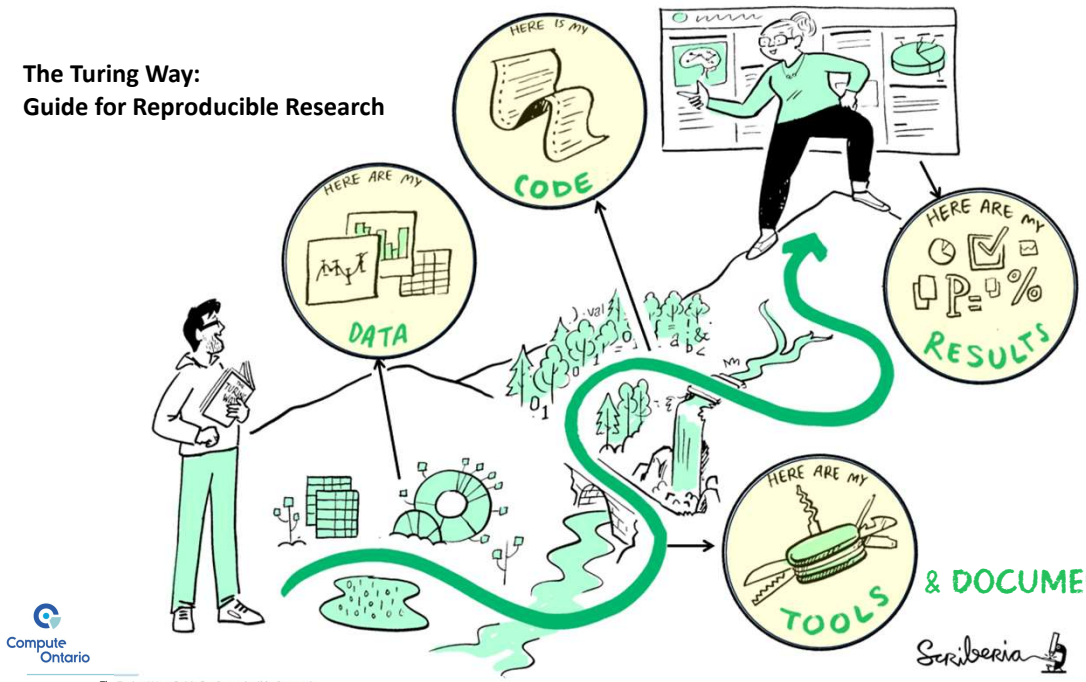
As a means of communicating scientific claims, computational reproducibility is imperative for **verifying and building upon findings**, for **preserving a complete scientific record**, and for **advancing pedagogy**. At present, this standard is rarely achieved.

33

Peer, L., Arguillas, F., Honeyman, T., Miljković, N., Peters-von-Gehlen, K., & CURE-FAIR WG Subgroup 3. (2021). Challenges of Curating for Reproducible and FAIR Research Output. *Research Data Alliance*. DOI: 10.15497/RDA00063 https://www.rd-alliance.org/sites/default/files/Challenges%20of%20Curating%20for%20Reproducible%20and%20FAIR%20Research%20Output%20-%20Output%20Card_0.pdf

33

The Turing Way:
Guide for Reproducible Research



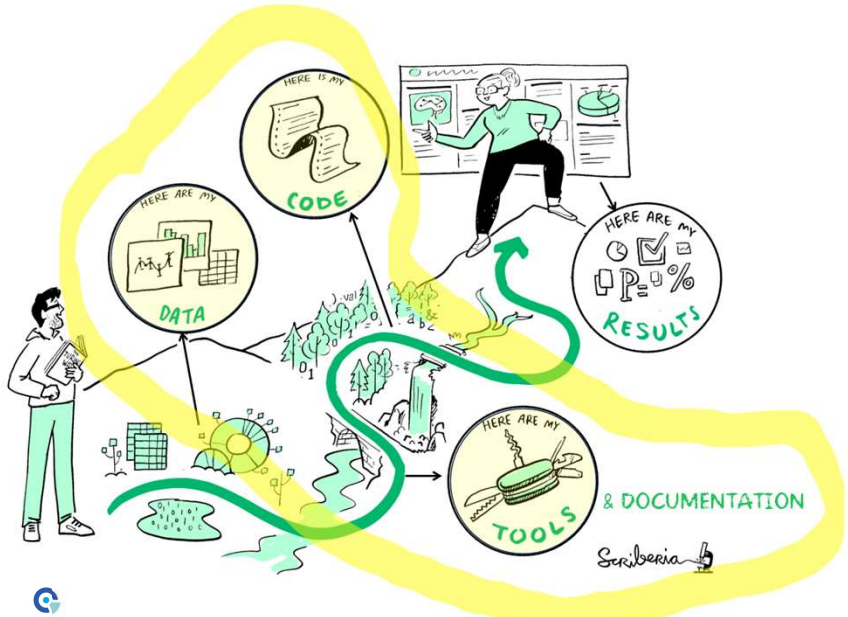
Compute Ontario

The Turing Way: Guide for Reproducible Research
https://the-turing-way.netlify.app/reproducible-research/reproducible-research.html

The Turing Way project illustration by Scriberia. Used under a CC-BY 4.0 licence. DOI: 10.5281/zenodo.3332807

34

34



"Reproducibility file bundle"

or

"***Research Compendium***"

CURE-FAIR

Compute Ontario

The Turing Way project illustration by Scriberia. Used under a CC-BY 4.0 licence. DOI: 10.5281/zenodo.3332807

10 Things for Curating Reproducible and FAIR Research [RDA Recommendation] Florio Arguillas, Thu-Mai Christian, Mandy Gooch, Tom Honeyman, Limor Peer, CURE-FAIR WG; 27 June 2022; **DOI**: 10.15497/RDA00074 or https://zenodo.org/records/6797657#.YuB6zcHMKrd

35

35

Challenges of Curating for Reproducible and FAIR Research Output

Findable

Accessible

Interoperable

Reusable

Compute Ontario

Challenges of Curating for Reproducible and FAIR Research Output - **A Report by the RDA CURE-FAIR Working Group, Subgroup 3 on CURE-FAIR Challenges**, Limor Peer, Florio Arguillas, Tom Honeyman, Nadica Miljković, Karsten Peters-von Gehlen and CURE-FAIR subgroup 3; April 12, 2021; https://zenodo.org/records/5094155#.YO0a8OgzaUk

36

https://www.nlm.nih.gov/oet/ed/cde/tutorial/02-300.html

36

Challenges of Curating for Reproducible and FAIR Research Output

Findable

- Difficulty finding data;
- Difficulty finding software;
- No software / data citation.

Accessible

Interoperable

Reusable

Compute Ontario

Challenges of Curating for Reproducible and FAIR Research Output - **A Report by the RDA CURE-FAIR Working Group, Subgroup 3 on CURE-FAIR Challenges**, Limor Peer, Florio Arguillas, Tom Honeyman, Nadica Miljković, Karsten Peters-von Gehlen and CURE-FAIR subgroup 3; April 12, 2021; https://zenodo.org/records/5094155#.YO0a8OgzaUk

37

https://www.nlm.nih.gov/oet/ed/cde/tutorial/02-300.html

37

Challenges of Curating for Reproducible and FAIR Research Output

Findable

Accessible

Interoperable

Reusable

- Difficulty finding data;
- Difficulty finding software;
- No software / data citation.

- Data, software, workflow, & digital objects not available due to:
  - proprietary software
  - high cost of archiving;
  - lack of a persistent identifier;
  - repository no longer exists;
  - dependencies and/or computing environment

Compute Ontario

Challenges of Curating for Reproducible and FAIR Research Output - **A Report by the RDA CURE-FAIR Working Group, Subgroup 3 on CURE-FAIR Challenges**, Limor Peer, Florio Arguillas, Tom Honeyman, Nadica Miljković, Karsten Peters-von Gehlen and CURE-FAIR subgroup 3; April 12, 2021; https://zenodo.org/records/5094155#.YO0a8OgzaUk

38

https://www.nlm.nih.gov/oet/ed/cde/tutorial/02-300.html

38

Challenges of Curating for Reproducible and FAIR Research Output

Findable

Accessible

Interoperable

Reusable

- Difficulty finding data;
- Difficulty finding software;
- No software / data citation.

- Data, software, workflow, & digital objects not available due to:
  - proprietary software
  - high cost of archiving;
  - lack of a persistent identifier;
  - repository no longer exists;
  - dependencies and/or computing environment

- Files don't work in another computing environment.

Compute Ontario

Challenges of Curating for Reproducible and FAIR Research Output - **A Report by the RDA CURE-FAIR Working Group, Subgroup 3 on CURE-FAIR Challenges**, Limor Peer, Florio Arguillas, Tom Honeyman, Nadica Miljković, Karsten Peters-von Gehlen and CURE-FAIR subgroup 3; April 12, 2021; https://zenodo.org/records/5094155#.YO0a8OgzaUk

39

https://www.nlm.nih.gov/oet/ed/cde/tutorial/02-300.html

39

Challenges of Curating for Reproducible and FAIR Research Output

Findable

Accessible

Interoperable

Reusable

- Difficulty finding data;
- Difficulty finding software;
- No software / data citation.

- Data, software, workflow, & digital objects not available due to:
  - proprietary software
  - high cost of archiving;
  - lack of a persistent identifier;
  - repository no longer exists;
  - dependencies and/or computing environment

- Files don't work in another computing environment.

- Little or no documentation;
- Code not working / not executable, or did not run as intended;
- Code obsolete or written in a different format;
- Incompatible software versions and/or operating systems;
- User licenses absent or unclear.

Compute Ontario

Challenges of Curating for Reproducible and FAIR Research Output - **A Report by the RDA CURE-FAIR Working Group, Subgroup 3 on CURE-FAIR Challenges**, Limor Peer, Florio Arguillas, Tom Honeyman, Nadica Miljković, Karsten Peters-von Gehlen and CURE-FAIR subgroup 3; April 12, 2021; https://zenodo.org/records/5094155#.YO0a8OgzaUk

40

https://www.nlm.nih.gov/oet/ed/cde/tutorial/02-300.html

40

Challenges of Curating for Reproducible and FAIR Research Output

Findable

Accessible

Interoperable

Reusable

- Difficulty finding data;
- Difficulty finding software;
- No software / data citation.

- Data, software, workflow, & digital objects not available due to:
  - proprietary software
  - high cost of archiving;
  - lack of a persistent identifier;
  - repository no longer exists;
  - dependencies and/or computing environment

- Files don't work in another computing environment.

- Little or no documentation;
- Code not working / not executable, or did not run as intended;
- Code obsolete or written in a different format;
- Incompatible software versions and/or operating systems;
- User licenses absent or unclear.

Compute Ontario

Challenges of Curating for Reproducible and FAIR Research Output - **A Report by the RDA CURE-FAIR Working Group, Subgroup 3 on CURE-FAIR Challenges**, Limor Peer, Florio Arguillas, Tom Honeyman, Nadica Miljković, Karsten Peters-von Gehlen and CURE-FAIR subgroup 3; April 12, 2021; https://zenodo.org/records/5094155#.YO0a8OgzaUk

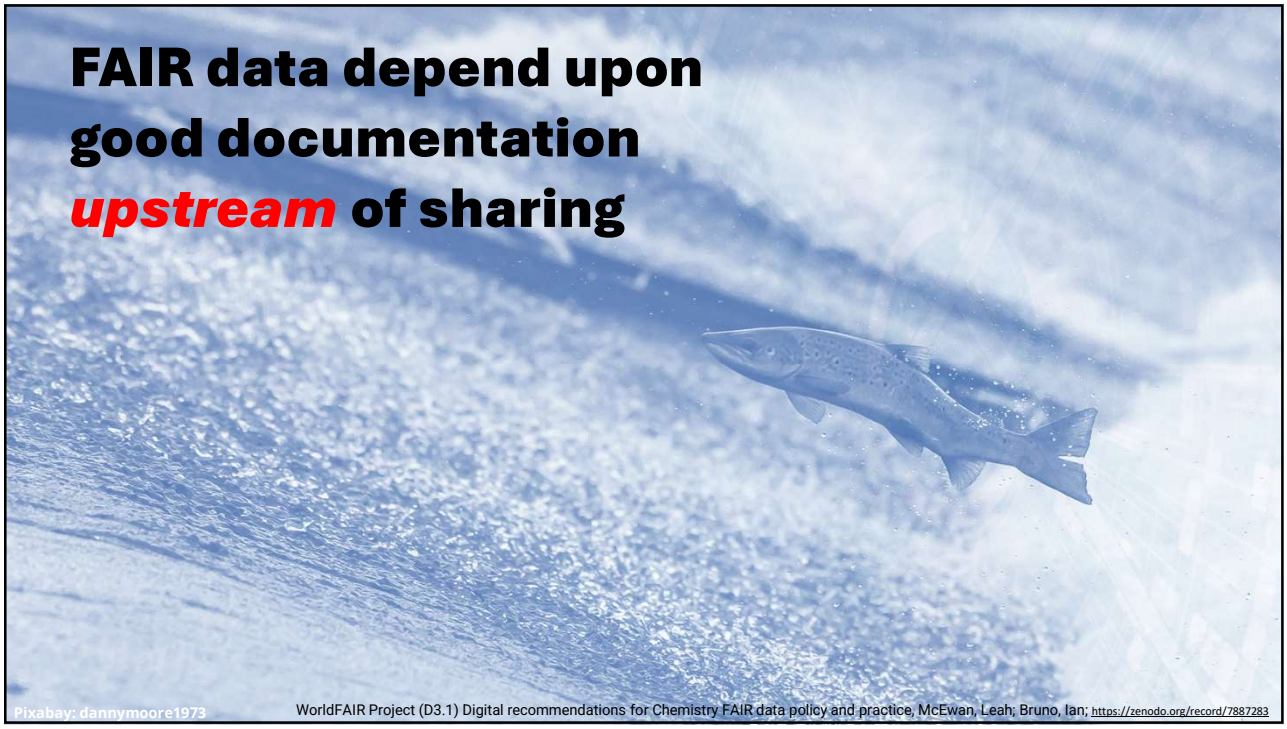
41

https://www.nlm.nih.gov/oet/ed/cde/tutorial/02-300.html

41


Pixabay: dannymoore1973

42


FAIR data depend upon
good documentation
upstream of sharing
Pixabay: dannymoore1973
WorldFAIR Project (D3.1) Digital recommendations for Chemistry FAIR data policy and practice, McEwan, Leah; Bruno, Ian; https://zenodo.org/record/7887283

43

'10 CURE-FAIR Things'

…will be of use to **data curators and information professionals**

Curators are "*often the first re-users of the research compendium*".

But should also "*be of interest to researchers, publishers, editors, reviewers, and others who have a stake in creating, using, sharing, publishing, or preserving reproducible research".*

Florio Arguillas, Thu-Mai Christian, Mandy Gooch, Tom Honeyman, Limor Peer, CURE-FAIR WG; 27 June 2022; **DOI**: 10.15497/RDA00074; https://zenodo.org/records/6797657#.YuB6zcHMKrd

44

Image by Tracy Lundgren from Pixabay

Have you included everything needed to reproduce your research in an organized and parsimonious way?

Thing 1: **Completeness**: Includes all data, metadata, and code needed to reproduce results.

Thing 2: **Organization**: Easy to understand and keep track of the various objects in the research compendium and their relationship over time.

Thing 3: **Economy**: Avoid extraneous objects in the compendium to minimize need for updates and/or maintenance over time.

Florio Arguillas, Thu-Mai Christian, Mandy Gooch, Tom Honeyman, Limor Peer, CURE-FAIR WG; 27 June 2022; **DOI**: 10.15497/RDA00074; https://zenodo.org/records/6797657#.YuB6zcHMKrd

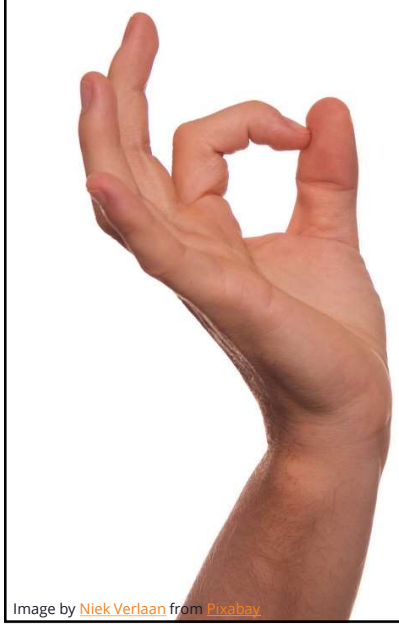
45

Is descriptive information about the research compendium and its components available and easy to understand?

Thing 4: **Transparency**: The research compendium provides full disclosure of the research process that produced the scientific claim.

Thing 5: **Documentation**: Information describing compendium objects is sufficiently detailed to enable independent understanding and use of the compendium.

46



Is information about the compendium and how it can be used available and easy to understand?

Thing 6: **Access**: Clear statement of who can use what, how, and under what conditions, with open access preferred.

Thing 7: **Provenance**: Origin and detailed versioning of compendium components provided.

47

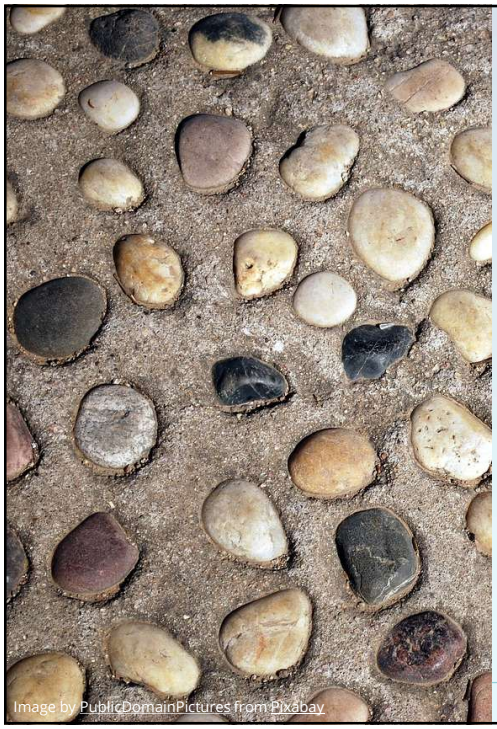


Image by PublicDomainPictures from Pixabay

Is information about the research compendium and its components embedded in code?

Thing 8: **Metadata**: Information about the research compendium and its components is embedded in a standardized, machine-readable code.

Thing 9: **Automation**: As much as possible, the computational workflow is script-based to facilitate re-execution using minimal actions.

48

Image by Bernd from Pixabay

Is there a plan for reviewing the research compendium for FAIR and computational reproducibility standards over time?

Thing 10: **Review**: A series of managed activities needed to ensure continued access to and functionality of the research compendium and its components for as long as necessary.

49

Thing 1: **Completeness**: Includes all data, metadata, and code needed to reproduce results.

Thing 2: **Organization**: Easy to understand and keep track of the various objects in the research compendium and their relationship over time.

Thing 3: **Economy**: Avoid extraneous objects in the compendium to minimize need for updates and/or maintenance over time.

Thing 4: **Transparency**: The research compendium provides full disclosure of the research process that produced the scientific claim.

Thing 5: **Documentation**: Information describing compendium objects is sufficiently detailed to enable independent understanding and use of the compendium.

Thing 6: **Access**: Clear statement of who can use what, how, and under what conditions, with open access preferred.

Thing 7: **Provenance**: Origin and detailed versioning of compendium components provided.

Thing 8: **Metadata**: Information about the research compendium and its components is embedded in a standardized, machine-readable code.

Thing 9: **Automation**: As much as possible, the computational workflow is script-based to facilitate re-execution using minimal actions.

Thing 10: **Review**: A series of managed activities needed to ensure continued access to and functionality of the research compendium and its components for as long as necessary.

50

10 Things for Curating Reproducible and FAIR Research [RDA Recommendation] Florio Arguillas, Thu-Mai Christian, Mandy Gooch, Tom Honeyman, Limor Peer, CURE-FAIR WG; 27 June 2022; **DOI**: 10.15497/RDA00074; https://zenodo.org/records/6797657#.YuB6zcHMKrd

50

Compute Ontario

51

https://www.pinterest.com/pin/41517627789600301/

51

"Scientific knowledge is cumulative. The production of each empirical finding should be viewed more as a **promissory note** than a final conclusion"

Compute Ontario

Image by Julita from Pixabay

Social, Behavioral, and Economic Sciences Perspectives on Robust and Reliable Science; Report of the Subcommittee on Replicability in Science Advisory Committee to the National Science Foundation Directorate for Social, Behavioral, and Economic Sciences, May 2015
https://www.nsf.gov/sbe/AC_Materials/SBE_Robust_and_Reliable_Research_Report.pdf

52

52

We should start viewing academic articles as **explicit promissory notes** that an **associated Research Compendium *exists*, *is available***, and ***is sufficient to support reproducibility***

"Scientific knowledge is cumulative. The production of each empirical finding should be viewed more as a **promissory note** than a final conclusion"

Compute Ontario

Image by Julita from Pixabay

Social, Behavioral, and Economic Sciences Perspectives on Robust and Reliable Science; Report of the Subcommittee on Replicability in Science Advisory Committee to the National Science Foundation Directorate for Social, Behavioral, and Economic Sciences, May 2015
https://www.nsf.gov/sbe/AC_Materials/SBE_Robust_and_Reliable_Research_Report.pdf

53

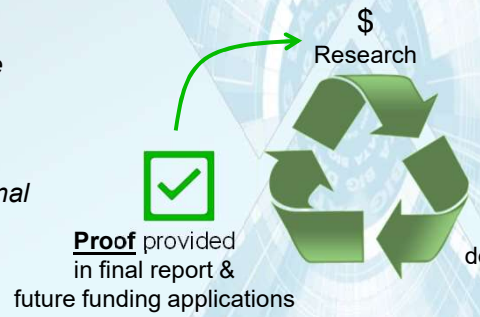
53

Reproducibility Recommendations from the US National Science Foundation

***Recommendation 1**: NSF-funded research must include detailed documentation to enable an independent researcher to reproduce the results of the original researcher.*

Proof of this must be provided in a project's Final Report and in future funding requests.

[paraphrased for brevity]



TRUST, BUT VERIFY

Compute Ontario

Image by PublicDomainPictures from Pixabay Image by Virgo Gem from Pixabay

Social, Behavioral, and Economic Sciences Perspectives on Robust and Reliable Science; Report of the Subcommittee on Replicability in Science Advisory Committee to the National Science Foundation Directorate for Social, Behavioral, and Economic Sciences, May 2015
https://www.nsf.gov/sbe/AC_Materials/SBE_Robust_and_Reliable_Research_Report.pdf

54

54

***Recommendation 2**: NSF should sponsor research that evaluates vari… whether a finding replicates and … under which circumstances … ing valid conclusions about replicability.*

Fund research on replicability

***Recommendation 3**: To permit assessing replication in various … to report associat… rics (e.g., standard… ct sizes, odds rat… with standard errors) and to assess the statistical significance of findings using these different methods.*

Encourage reporting of different metrics to help assess statistical significance

***Recommendation 4**: NSF should sponsor research that identifie… ng all types of generali… participants to a pop… measures, from one set of circumstances to other circumstances) and diff… icate.*

Fund research on generalizability of findings

***Rec… ing the opti… results so a…*

Fund research on optimal and minimum statistical reporting standards to facilitate meta-analyses

***Recommendation 6**: NSF should support research into the use of questionable research practices, the causes that encourage su… entions int… rt the identification of empirically validated optimal research practices to avoid the production of illusory findings.*

Fund research on bad research behaviour(s) and how to address them

***Recommendation 7**: In NSF grant proposals, investigators should be required to describe plans for implementing and fully reporting test… al met… uld be enc… the… eriously con… reports and final reports, investigators should be required to describe whether more than one hypothesis was considered, the robustness checks conducted and results obtained.*

Require grant applicants to fully describe statistical approaches, alternate analytical approaches, and other hypotheses considered

***Recommendation 8**: NSF should sponsor research seeking … non-robust research findings.*

Fund research to document '*suboptimal practices*' to call them out and effect change

***Recommendation 9**: NSF should create a Foundation-wide committee of experts to monitor issues of reproducibility, … Foundation, to propose ways to change the NSF granting process to enhance scientific quality and efficiency, and to provide leadership on these issues in the coming decades.*

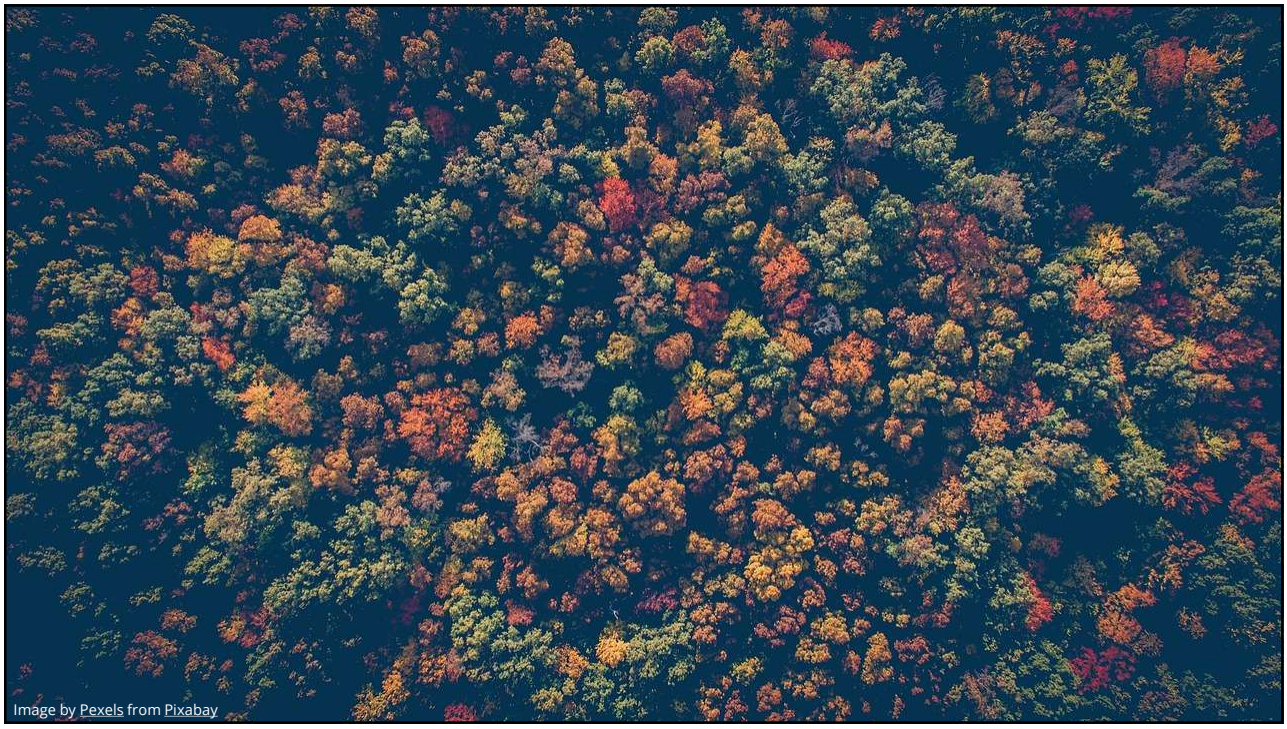
Create an NSF-wide expert committee to monitor and address issues of reproducibility

Social, Behavioral, and Economic Sciences Perspectives on Robust and Reliable Science; Report of the Subcommittee on Replicability in Science Advisory Committee to the National Science Foundation Directorate for Social, Behavioral, and Economic Sciences, May 2015
https://www.nsf.gov/sbe/AC_Materials/SBE_Robust_and_Reliable_Research_Report.pdf

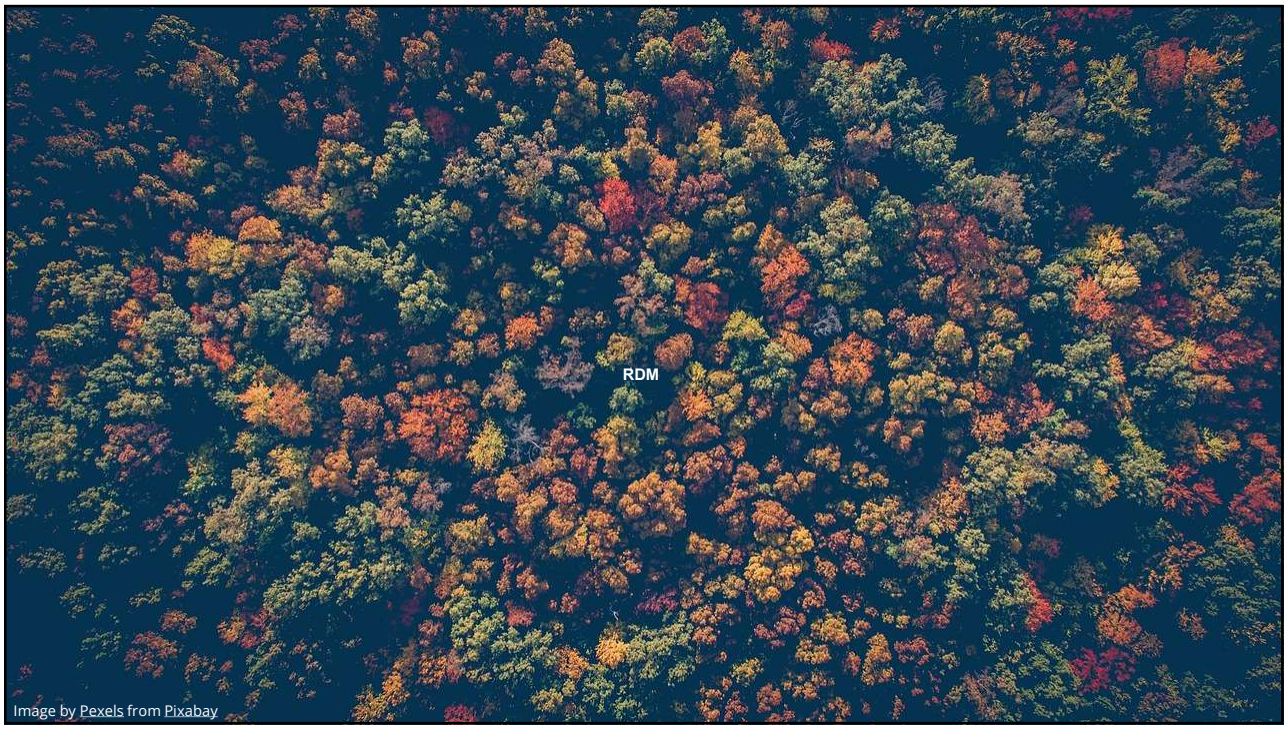
55

55

56



57



58

59

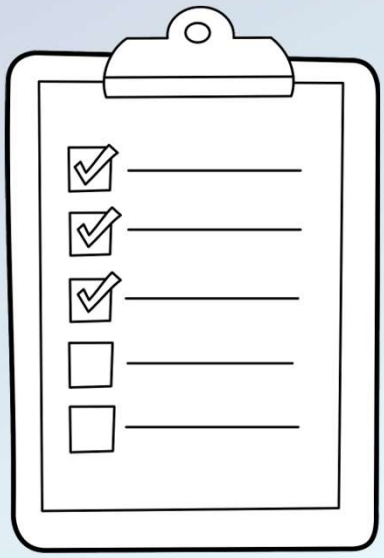
Curating Data Sets for Reproducibility Workshop

Qian Zhang (U. Waterloo)
Sandra Sawchuk (Mount Saint Vincent U.)
Shahira Khair (U. Victoria)

https://research-reuse.github.io

Compute Ontario

60



Reproducibility Checklists

Du, X., Aristizabal-Henao, *et al,* 2022. **A Checklist for Reproducible Computational Analysis in Clinical Metabolomics Research**. Metabolites, 12(1), 87–. https://doi.org/10.3390/metabo12010087

Hutton, C., *et al,* 2016. **Most computational hydrology is not reproducible, so is it really science?** Water Resources Research, 52(10), 7548–7555. https://doi.org/10.1002/2016WR019285

Sandve, Geir Kjetil, *et al*. **Ten Simple Rules for Reproducible Computational Research.** PLOS Computational Biology 9 , no. 10 (2013): 1-4.

Software as research data: Tips for Reproducible Research based on Software. https://zenodo.org/record/2611314/files/Software-Manage-Publish-EN.pptx?download=1

British Ecological Society. A Guide to Reproducible Code in Ecology and Evolution, 2017. https://www.britishecologicalsociety.org/wp-content/uploads/2017/12/guide-to-reproducible-code.pdf

Joelle Pineau, 2020. **The Machine Learning Reproducibility Checklist**. https://www.cs.mcgill.ca/~jpineau/ReproducibilityChecklist.pdf

James, Wilkins-Diehr, *et al*, (2014). **Standing Together for Reproducibility in Large-Scale Computing: Report on reproducibility@XSEDE**. https://doi.org/10.48550/arxiv.1412.5557

Crick, Hall, B. A., & Ishtiaq, S. (2017). **Reproducibility in Research: Systems, Infrastructure, Culture**. Journal of Open Research Software, 5(1), 32–. https://doi.org/10.5334/jors.73

Compute Ontario

61

61

Questions?

Jeff Moon
moonj@computeontario.ca

62

63

63